RESEARCH ARTICLE

Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome

Kayoko Yamada,[1]* Jun Lim,[2]* Joseph M. Dale,[1] Huaming Chen,[2,3] Paul Shinn,[2,3] Curtis J. Palm,[4] Audrey M. Southwick,[4] Hank C. Wu,[1] Christopher Kim,[2,3] Michelle Nguyen,[4] Paul Pham,[1] Rosa Cheuk,[2,3] George Karlin-Newmann,[4] Shirley X. Liu,[1] Bao Lam,[4] Hitomi Sakano,[1] Troy Wu,[4] Guixia Yu,[1] Molly Miranda,[4] Hong L. Quach,[1] Matthew Tripp,[4] Charlie H. Chang,[1] Jeong M. Lee,[1] Mitsue Toriumi,[1] Marie M. H. Chan,[1] Carolyn C. Tang,[1] Courtney S. Onodera,[1] Justine M. Deng,[1] Kenji Akiyama,[5] Yasser Ansari,[2] Takahiro Arakawa,[6] Jenny Banh,[1] Fumika Banno,[1] Leah Bowser,[4] Shelise Brooks,[3] Piero Carninci,[6,7] Qimin Chao,[3] Nathan Choy,[2] Akiko Enju,[5] Andrew D. Goldsmith,[1] Mani Gurjal,[4] Nancy F. Hansen,[4] Yoshihide Hayashizaki,[6,7] Chanda Johnson-Hopson,[3] Vickie W. Hsuan,[1] Kei Iida,[5] Meagan Karnes,[2] Shehnaz Khan,[3] Eric Koesema,[2] Junko Ishida,[5] Paul X. Jiang,[1] Ted Jones,[4] Jun Kawai,[6,7] Asako Kamiya,[5] Cristina Meyers,[2] Maiko Nakajima,[5] Mari Narusaka,[5] Motoaki Seki,[5,8] Tetsuya Sakurai,[5] Masakazu Satou,[5] Racquel Tamse,[4] Maria Vaysberg,[1] Erika K. Wallender,[1] Cecilia Wong,[1] Yuki Yamamura,[1] Shiaulou Yuan,[1] Kazuo Shinozaki,[5,8] Ronald W. Davis,[4,9] Athanasios Theologis,[1]*† Joseph R. Ecker[2,3]*†

Functional analysis of a genome requires accurate gene structure information and a complete gene inventory. A dual experimental strategy was used to verify and correct the initial genome sequence annotation of the reference plant *Arabidopsis*. Sequencing full-length cDNAs and hybridizations using RNA populations from various tissues to a set of high-density oligonucleotide arrays spanning the entire genome allowed the accurate annotation of thousands of gene structures. We identified 5817 novel transcription units, including a substantial amount of antisense gene transcription, and 40 genes within the genetically defined centromeres. This approach resulted in completion of ~30% of the *Arabidopsis* ORFeome as a resource for global functional experimentation of the plant proteome.

The genome sequence of *Arabidopsis thaliana* serves as a reference for plants (*1*). The initial identification of transcriptional units in the *Arabidopsis* genome sequence was carried out largely by ab initio gene predictions, sequence homology, sequence motif analysis, and other nonexperimental methods (*2*–*7*). This led to an estimate of 25,500 protein-coding genes (*2*). Though gene prediction software has steadily improved, the ability of these programs to precisely determine gene structures in sequenced genomes remains unsatisfactory (*8*–*10*). A recent attempt to verify experimentally the accuracy of the *Arabidopsis* genome annotation with conventional molecular approaches proved quite inefficient in identifying full-length (fl) open-reading frames (ORFs) (*11*). Global experimental approaches are expected to greatly improve genome annotation (*12*, *13*).

A genome sequence that is empirically annotated can provide the foundation for the determination of its ORFeome, the complete set of ORF clones for all protein-coding genes. Access to a "gold standard" cDNA/ORF clone collection (*14*–*16*), representing the entire *Arabidopsis* proteome, is urgently needed as a common resource for research (*17*). Here we report the experimental definition of the transcriptional units for all *Arabidopsis* genes by fl-cDNA discovery and by hybridization of RNA populations to whole-genome arrays (WGAs) (fig. S1).

Annotation and fl cDNAs. Dramatic improvements in genome annotation have been achieved for several organisms by means of sequences of fl-cDNAs (*8*, *15*, *18*, *19*). For *Arabidopsis*, three collections of fl-cDNAs have been used for this purpose (*8*, *20*–*23*). Of the 26,828 predicted genes in the *Arabidopsis* genome, 25,540 were annotated as protein-coding, with the rest being annotated as pseudo- and partial genes (fig. S3). A unique location in the genome can be identified for the majority (99%) of cDNAs (fig. S4). From this analysis, three classes of genes were detected in the annotated genome (fig. S5): (i) annotated expressed (AE) genes; (ii) annotated non-expressed (ANE) genes, and (iii) newly discovered genes located in the intergenic regions (non-annotated expressed; NAE). Seventy percent (18,093) of the 25,540 annotated genes were found to be transcriptionally active on the basis of identification of an expressed sequence tag (EST) or cDNA, and these represent the class of the AE genes. The remaining 7447 genes (30%) are the hypothetical, ANE (*24*) genes. Only 10,507 (58%) of the AE genes have at least one fl-cDNA clone (Fig. 1). The remaining 7586 genes (42%) have evidence of their expression from At-ESTs and/or non–fl-cDNAs (Fig. 1). Among the 10,507 genes with a fl-cDNA sequence, the annotated structures of 32% of them (3336) were inaccurately annotated. Incorporation of fl-cDNA sequences resulted in the dramatic improvement of the genome annotation (fig. S6). The vast majority of predicted gene structures in the initial release of the *Arabidopsis* genome annotation (August 2001 V1.0) were not derived from experimental evidence of transcription.

Genome annotation, intergenic regions, and comparative genomics. Alignment of the various At-ESTs and fl-cDNAs on the V1.0 genome annotation reveals that a considerable number of cDNAs mapped within in-

[1]Plant Gene Expression Center, Albany, CA 94710, USA. [2]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [3]The *Arabidopsis thaliana* Genome Center, Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. [4]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA. [5]Plant Mutation Exploration Team, Plant Functional Genomics Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Tsukuba 305-0074, Japan. [6]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Main Campus, Wako 351-0198, Japan. [7]Laboratory for Genome Exploration Research Group, RIKEN GSC, Yokohama Institute, Kanagawa 230-0045, Japan. [8]Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, Tsukuba 305-0074, Japan. [9]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA.

*These authors contributed equally to this manuscript.
†To whom correspondence should be addressed. E-mail: theo@nature.berkeley.edu (A.T.) or ecker@salk.edu (J.R.E.)

tergenic regions (IGRs) (fig. S5). Clustering analysis (*22*) reveals that as many as 1347 NAE genes may reside in the IGRs. The density and distribution of the new unpredicted genes are shown in fig. S7 (*22*). The availability of a substantial number of *Brassica*, rice, and wheat ESTs allowed us to examine whether any of the aANE genes (Fig. 1) show similarity to these non-*Arabidopsis* ESTs. This analysis suggests that 45% of the ANE *Arabidopsis* genes are transcriptionally active [BLAST e-value $< 10^{-25}$, (*22*)] (Fig. 1A). Furthermore, alignment of the *Brassica*, rice, or wheat ESTs within the IGRs identified an additional 1042 transcriptionally active regions (BLAST e-value $< 10^{-25}$, Fig. 1B). Overall, our analysis revealed that the *Arabidopsis* genome contains ~28,000 genes of which 15,033 (59%) annotated genes lack a fl-cDNA clone and, therefore, have not been experimentally verified.

Mapping transcriptional units using high-density oligonucleotide tiling arrays. In the absence of a corresponding fl-cDNA for 15,033 annotated genes (7586 AE and 7447 ANE genes) and for the ~1350 newly discovered genes located in the IGRs (Fig. 1B), new methods are needed to obtain evidence for the existence of this remaining large number of "untouched" genes. Recently, microarray technology has been used for mapping transcription units in genomes (*12*, *13*, *25*). Despite the fact that the *E. coli* genome is well characterized, nearly 25% of the transcripts were previously unidentified (*26*). Except for *Saccharomyces cerevisiae* (*27*, *28*), however, tiling arrays have not been applied on a genome-wide scale for unbiased examination of transcriptional activity in an eukaryote.

To identify transcription units in the *Arabidopsis* genome, we used custom high-density oligonucleotide arrays that tile the entire genome and RNA samples prepared from a diverse set of tissues and treatments to ensure broad representation of transcriptional activity. Four pilot arrays, identical in design except for the regions of the genome sequence covered, were used to test the feasibility of the approach (fig. S8) (*22*). In brief, labeled cRNAs were prepared from different plant tissues and hybridized to the four arrays. Figure 2A shows a deconvoluted and merged image of these arrays after hybridization with cRNA from three different plant tissues (flower, etiolated seedlings, and inflorescence). Differential gene expression was revealed across the 5 Mb region. For example, gene At1g12110 (Fig. 2B) is expressed only in seedlings; gene At1g10770 (Fig. 2C), only in flowers; and gene At1g10630, in all three tissues (Fig. 2D). In addition, the hybridization profiles revealed the structures of the individual genes as confirmed by the sequences of fl-cDNAs corresponding to each of these genes (Fig. 2, B to D). Further analysis with mRNAs isolated from light grown (L+), dark grown (L–) and cold-stressed light grown (L+ cold) seedlings (*22*) representing different growth conditions revealed that 950 (88%) of the AE transcription units present on the four chips could be detected (Fig. 3A, AE). One-half of the ANE transcription units present on these four chips were also detected (Fig. 3A, ANE). Expression was detected for 61% percent of the newly discovered genes present on the four pilot chips (Fig. 3A, NAE).

Validation of tiling chip–detected transcription units was carried out through reverse transcriptase–polymerase chain reaction (RT-PCR) amplification, cloning, and sequencing (fig. S9) (*22*). RT-PCR products were detected for 78% of the tested transcriptional units (table S1). Fifty-seven percent of the RT-PCR products yielded "perfect" ORF clones from both AE and ANE genes (table S1). The remaining 42% of the clones were without ORFs due to stop codons introduced during RT-PCR amplification (table S1).

Whole-genome arrays. We next designed a set of 12 oligonucleotide arrays representing ~94% of the *Arabidopsis* genome sequence [110 Mb (*22*) (fig. S10)]. Each array contains ~834,000 25-mer oligos (*22*). Four RNA populations (*22*) were hybridized to these arrays, and a transcription map for the entire *Arabidopsis* genome was determined (figs. S11 to S15) (*22*). Expression of 84% of the AE genes and 37% of the ANE genes was detected when these four RNA samples were used (Fig. 3B, AE and ANE). In addition, we were also able to detect a large number (~54%) of transcriptionally active sites corresponding to the chromosomal locations where the newly discovered genes were found (Fig. 3B, NAE). Transcriptional activity was also detected in 2000 intergenic regions (23%) that were thought to be devoid of transcriptional activity (Fig. 3B, IGR).

The five chromosomes were equally transcriptionally active under various growth conditions or developmental stages (fig. S16).

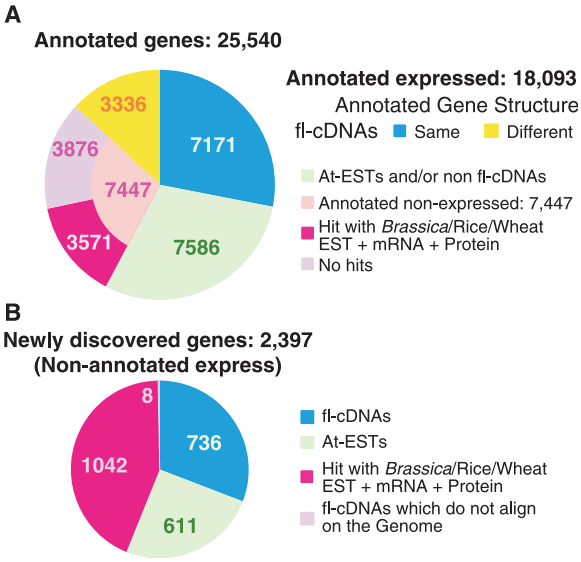

Fig. 1. Quality of the *Arabidopsis* genome annotation V1.0. (**A**) AE and ANE genes (hypothetical). (**B**) Newly discovered genes. The fl-cDNAs include: community full length (CFL); Ceres (*8*); Riken *Arabidopsis* full length [RAFL, (*20*)] and Chip (C) clones isolated during this study. See (*46*) for percentage distribution in the two types of annotated genes.

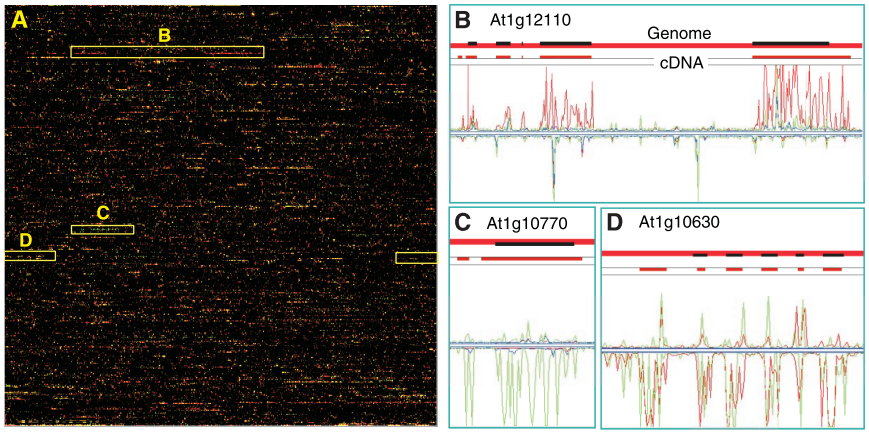

Fig. 2. Pilot Affymetrix tiling genome arrays for mapping transcription units. (**A**) Deconvoluted hybridization image of array 1A hybridized with three different mRNAs from seedlings, flowers, and young inflorescence. The chip images from three independent hybridizations were overlaid electronically and shown on (A). (**B**) Detection of a transcription unit specifically expressed in a 3-day-old etiolated seedlings treated with ethylene. (**C**) Detection of a transcription unit specifically expressed in flowers. (**D**) Detection of a transcription unit expressed in all three tissues.

Transcriptional activity across the chromosomes using four different messenger RNAs (mRNAs) is quantitatively distinct (Fig. 4). The transcriptional activity across each chromosome was different in various tissues tested (Fig. 4). For example, the right arm of chromosome 2 showed more transcriptionally active sites in suspension cell cultures [Fig. 4, compare (E) with (C) in chr. 2] than in flowers. In addition, in certain cases the transcriptional activity was strand-specific. For example, in roots there was more transcription from the forward-strand than from the reverse complement strand (Fig. 4D, chr. 1). The gene density is approximately the same in both strands. The distribution of tissue-specific transcripts across the five chromosomes is shown in Fig. 5. Functional classification of the tissue-specific transcripts revealed variation in the types of genes expressed in a tissue-specific manner (fig. S17). For example, 16% of the flower-specific AE transcription units encode proteins involved in metabolism, whereas only 6.5% of the same class of genes were detected in suspension cell cultures (fig. S17C, compare flower with suspension cell culture).

More detailed analyses of the whole-genome transcription map focused on examination of regions of the genome, such as pseudogenes and centromeres, thought to be relatively transcriptionally silent and also on genomic locations showing antisense RNA expression. (Fig. 6A). We detected transcription for ~20% of the 1332 ANE "pseudogenes" (Fig. 6A), suggesting that these ORFs may be misannotated. Alternatively, as recently described for mouse (*29*), such actively transcribed "pseudogenes" may function to regulate gene activity. In total, ~7600 annotated genes (~30% of all annotated genes) were identified that showed significant antisense RNA expression (Fig. 6B), suggesting that double-strand RNA formation may be a general phenomenon in plant cells (*30*). We found complementary patterns of tissue-specific expression of sense and antisense RNAs for many genes, indicating a possible biological role for these transcripts (*31*). Lastly, we surveyed regions located within the genetically defined centromeres for transcriptional activity [(*2*, *32*); Fig. 6C]. Transcriptional activity was confirmed for ~90% of the AE genes. Transcriptional activity was also observed for ~45% of the "unexpressed" genes, and we identified 29 new transcribed genes within the five centromere regions. Surprisingly, two unusually highly expressed sites, "hot spots," were identified within the genetically defined centromeres of two chromosomes (*2*, *22*, *32*) (Fig. 6D). The centromere region of chromosome 2 (CEN2) contains a large (~600 kb) insertion of a rearranged copy of the mitochondrial genome (*2*). This entire genomic location showed high level of transcriptional activity. In addition, several unannotated regions within the centromere region of chromosome 3 (CEN3) were identified that displayed a significantly higher level of transcriptional activity from both forward and reverse complement DNA strands (Fig. 6D). These highly transcribed sequences within CEN3 were largely composed of transposon or retrotransposon-like sequences but also contained some unique sequences.

Overall, evidence from both cDNA sequencing and WGA expression studies allowed detection of transcriptional activity for 5817 previously ANE (hypothetical) genes (Fig. 3B, ANE; figs. S6, A and C).

Determination of the *Arabidopsis* ORFeome. Large-scale ORFeome projects not only aid in the improvement of whole genome annotation, but also provide valuable resources to the community for functional genomic studies and proteomics (*33*, *34*). We constructed 8750 error-free ORF clones [Table 1 (*22*, *35*)] representing approximately 32% of the annotated genes. Functional classification of the constructed ORFeome revealed a similar distribution of the various functional classes compared with the complete set of annotated genes (fig. S19). Figure S20 shows the distribution of the ORF clones constructed among the various available fl-cDNAs.

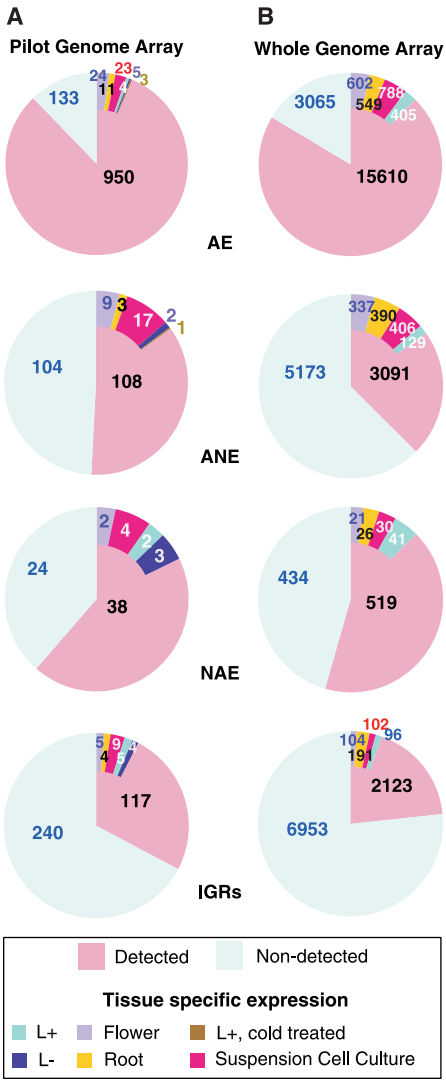

Fig. 3. Quantification of total and tissue-specific transcriptional units detected by the pilot (**A**) and whole genome (**B**) arrays of AE genes, ANE genes, NAE genes, and IGRs. See (*22*) for additional details. The number of genes detected for each category present in the arrays are as follows: (A) AE, 950/1083 (88%); ANE, 108/212 (51%); NAE, 38/62 (61%); IGRs, 117/357 (33%). (B) AE, 15,610/18,675 (84%); ANE, 3091/8264 (37%); NAE, 519/953 (54%); IGRs, 2123/9076 (23%).

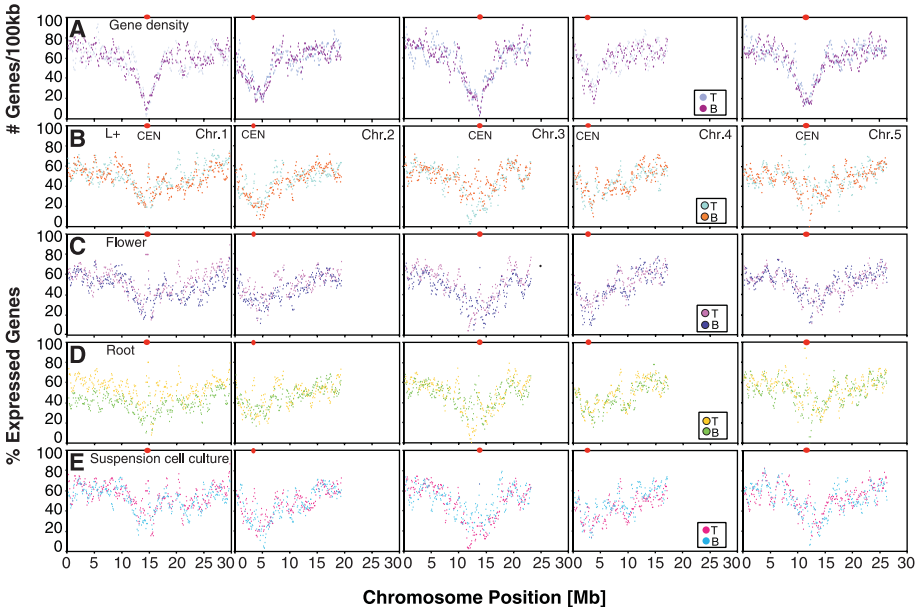

Fig. 4. Global gene expression detected by whole genome arrays in the *Arabidopsis* chromosomes. (**A**) Predicted gene density on each strand of the five chromosomes. (**B**) Percentage of expressed genes on each strand of the five chromosomes in light-grown seedlings. (**C**) Flowers. (**D**) Roots. (**E**) Suspension cell culture. Forward strand is indicated as top (T), whereas reverse strand as bottom (B). The genetically defined centromeres were indicated as red dots.

Utility of WGAs for genomic analysis. We used a strategy for mapping transcription units that included both a traditional cDNA cloning and DNA sequencing approach and a novel high-throughput approach that utilizes tiling array technology (*12*, *13*, *25*). Sequencing ESTs and fl-cDNAs verified 40% of the annotated genes, dramatically improved genome annotation, and allowed the construction of 30% of the ORFeome. As with other large-scale cDNA-based gene collection projects (*8*, *15*, *18*), the traditional approach of sequencing fl-cDNAs reaches a point of diminishing returns: two-thirds of the total annotated *Arabidopsis* genes still have no corresponding fl-cDNA. The WGA technology allowed us to identify transcripts for the ANE genes and enabled construction of the remaining 30% of the *Arabidopsis* ORFeome. We were able to detect ~60% of the total annotated genes (AE and ANE) using only a limited number of RNA samples (four). A more comprehensive survey of tissues and cell types should enable a more complete description of the transcriptome.

WGAs serve as a "molecular roadmap" for detecting transcription units and their structures across the chromosomes. This capability arrives at a convenient time, because several major efforts to generate fl-cDNAs for identification of a full set of ORF clones of expressed genes are under way [the Mammalian Gene Collection (*15*, *16*) and FANTOM (*36*)]. As these collections become more complete, the marginal utility of each added EST decreases, and the prospect for discovering new genes diminishes. Use of full-genome tiling arrays is a natural way to complete the identification of transcription units. In addition, the data presented here raise the prospect that novel, tissue-specific transcription units can be identified in *Arabidopsis* by means of WGAs, cloned, and functionally studied with biochemical and reverse genetic approaches (*37*). The WGAs allowed us to detect transcription from many of the "pseudogenes," providing additional evidence that this class of genes may function in the regulation of their paralogous counterparts (*29*). Our studies also revealed that a large fraction of *Arabidopsis* genes showed expression of an antisense RNA transcript that, in many cases, was tissue-specific. These results suggest that the production of double-stranded RNAs may play a greater role in the regulation of normal plant development than is currently understood.

The use of genome tiling array technology is not restricted to transcription unit mapping of RNA polymerase II transcripts. WGAs can also be used for detecting mitochondrial- or plastid-derived RNAs and potentially could be used for the identification of a variety of small noncoding RNAs (nc-RNAs) (*38*). By design, WGAs contain DNA sequences absent from traditional gene expression arrays, such as the promoters, introns, and intergenic "dark matter." Thus, they are well suited for global mapping of DNA binding sites by location analysis (*39*, *40*) and provide an "unbiased view" of DNA binding sites that recent studies suggest may lie within coding regions at a greater frequency than previously suspected (*41*). Genome tiling arrays can

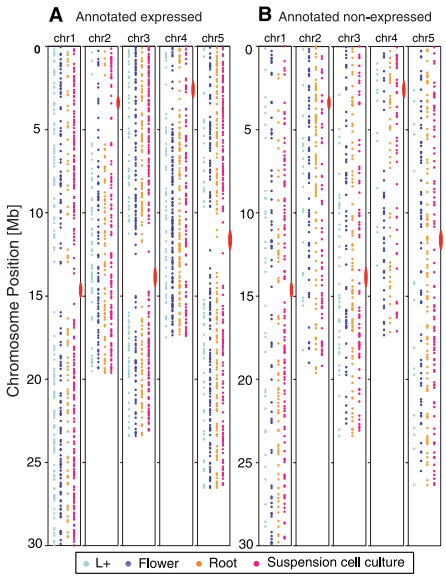

Fig. 5. Chromosomal locations of tissue-specific expressed transcription units. (**A**) AE genes. (**B**) ANE genes.

Table 1. ORF clones constructed.

Class of genes	Type of ORF clone*	
	U	C
Annotated expressed	7685	486
Annotated non-expressed	N/A	275
Newly discovered	171	27
Annotated as pseudogenes	19	87
Total	7875	875

*See (*22*) for clone terminology.

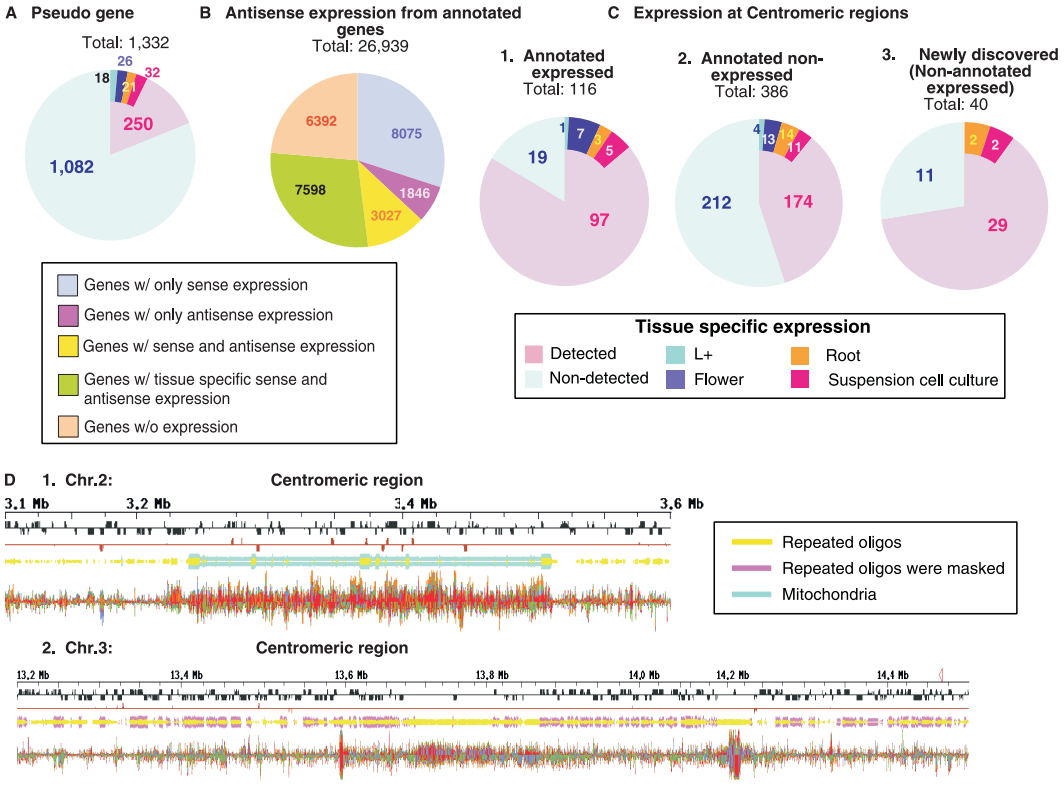

Fig. 6. WGA analysis of transcription of pseudogenes, antisense RNAs, and centromere-located genes. Proportion of AE and ANE genes for (**A**) predicted pseudogenes, (**B**) antisense genes, or (**C**) centromere regions. (**D**) Transcriptionally active "hot spots" with the genetically defined centeromeres for chromosomes 2 and 3. See SOM text for further description.

also be used to detect DNA sequence polymorphisms and potentially to detect differences in cytosine methylation patterns among *Arabidopsis* ecotypes, which would be useful for rapid genome-wide mapping of quantitative trait loci (QTL) (*42*). Finally, WGAs might also be used for direct mapping of point mutations without the use of segregated populations.

The impact of these findings and of the experimental resources developed is not restricted to *Arabidopsis*. For example, 85% of *Arabidopsis* genes have close homologs in the rice genome (*43*). Therefore, this resource will aid in the elucidation of the function(s) of the vast majority of genes in plant genomes (*43*, *44*), providing fundamental knowledge that we hope will eventually lead to the engineering of new plant species.

References and Notes

1. E. M. Meyerowitz, in *Arabidopsis*, E. M. Meyerowitz and C. Somerville, Eds. (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1994).
2. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
3. X. Lin *et al.*, *Nature* **402**, 761 (1999).
4. K. Mayer *et al.*, *Nature* **402**, 769 (1999).
5. M. Salanoubat *et al.*, *Nature* **408**, 820 (2000).
6. S. Tabata *et al.*, *Nature* **408**, 823 (2000).
7. A. Theologis *et al.*, *Nature* **408**, 816 (2000).
8. B. J. Haas *et al.*, *Genome Biol.* **3**, 1 (2002).
9. C. Mathe, M.-F. Sagot, T. Schiex, P. Rouze, *Nucleic Acids Res.* **30**, 4103 (2002).
10. M. Q. Zhang, *Nature Rev. Genet.* **3**, 698 (2002).
11. Y. L. Xiao, M. Malik, C. A. Whitelaw, C. D. Town, *Plant Physiol.* **130**, 2118 (2002).
12. D. D. Shoemaker *et al.*, *Nature* **409**, 922 (2001).
13. J. L. Rinn *et al.*, *Genes Dev.* **17**, 529 (2003).
14. L. Brizuela, A. Richardson, G. Marsischky, J. Labaer, *Arch. Med. Res.* **33**, 318 (2002).
15. Mammalian Gene Collection (MGC) Program Team, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 16899 (2002).
16. R. L. Strausberg, E. A. Feingold, R. D. Klausner, F. S. Collins, *Science* **286**, 455 (1999).
17. National Academy of Sciences Report. *National Plant Genome Initiative: Objectives for 2003-2008* (National Academies Press, Washington, DC, 2002). Available at: www.nap.edu/books/0309085217/html.
18. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
19. N. Osato *et al.*, *Genome Res.* **12**, 1127 (2002).
20. M. Seki *et al.*, *Science* **296**, 141 (2002).
21. A large collection of ~20,150 RIKEN *Arabidopsis* full-length (RAFL)-cDNAs were produced by the RIKEN Genome Sciences Center (*21*) and the complete sequence of the majority of RAFLs was determined in this study [fig. S2; (*22*)].
22. Materials and Methods are available as supporting material on *Science* Online.
23. A collection of ~2300 fl-cDNAs have been constructed by various laboratories known as the CFLs (community full length) whose construction preceded the two large fl-cDNA collections (*22*).
24. J. Reboul *et al.*, *Nature Genet.* **27**, 332 (2001).
25. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
26. B. Tjaden *et al.*, *Nucleic Acids Res.* **30**, 3732 (2002).
27. E. A. Winzeler *et al.*, *Science* **281**, 1194 (1998).
28. L. M. Steinmetz, R. W. Davis, *Biotechnol. Genet. Eng. Rev.* **17**, 109 (2000).
29. S. Hirotsune *et al.*, *Nature* **423**, 91 (2003).
30. C. Llave, K. D. Kasschau, M. A. Rector, J. C. Carrington, *Plant Cell* **14**, 1605 (2002).
31. R. Yelin *et al.*, *Nature Biotechnol.* **21**, 379 (2003).
32. G. P. Copenhaver *et al.*, *Science* **286**, 2468 (1999).
33. M. Vidal, *Cell* **104**, 333 (2001).
34. P. Carninci et al., *Genomics* **77**, 79 (2001).
35. The majority of ORF clones (7875) were constructed by transferring the ORFs from the RAFL clones into the pUNI51 cloning vector (*45*). The remaining 875 ORFs consisted of 594 chip-derived RT-PCR clones produced to replace defective RAFL clones and ORF clones for 282 new ANE genes. The ORF PCR products were subcloned as *SfiA*/*SfiB* fragments (fig. S18), allowing unidirectionality of the cloning process.
36. FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team, *Nature* **420**, 563 (2002).
37. J. Alonso *et al.*, *Science* **301**, 653 (2003).
38. S. R. Eddy, *Nature Rev. Genet.* **2**, 919 (2001).
39. B. Ren *et al.*, *Science* **290**, 2306 (2000).
40. V. R. Iyer *et al.*, *Nature* **409**, 533 (2001).
41. A. Morillon, J. O'Sullivan, A. Azad, N. Proudfoot, J. Mellor, *Science* **300**, 492 (2003).
42. J. O. Borevitz *et al.*, *Genome Res.* **13**, 513 (2003).
43. S. A. Goff *et al.*, *Science* **296**, 92 (2002).
44. M. Kirst *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7383 (2003).
45. Q. H. Liu, M. Z. Li, D. Leibham, D. Cortez, S. J. Elledge, *Curr. Biol.* **8**, 1300 (1998).
46. The distribution of the various types of fl-cDNAs that correspond to genes with the same or different annotated gene structure is as follows: same annotated structure, CFL, 20%; Ceres, 33%; RAFL, 43%; C-clones, 3%; different annotated structure, CFL, 10%; Ceres, 27%; RAFL, 61%.

47. We thank M. Johnston for critical reading of the manuscript, K. Mayer for gene functional category information, and T. Gingeras for support of this project. We also thank S. Elledge for providing us with the pUNI50 vector and P. Surko, J. Borevitz, and T. Mockler for useful discussions. Supported by the NSF Plant Genome Research Program under awards DBI-9975718, DBI-0196098 (to J.R.E.), DBI-9872752 (to R.W.D.), and USDA CRIS no. 5335-21430-005-00D (to A.T.). This study has also been supported by Research Grant for Genome Research from RIKEN (to K.S.) and by Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (to Y.H.). Affymetrix genome tiling array expression data have been deposited in the Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo/). Accession numbers are as follows: pilot tiling arrays, GSM8999 through GSM9010 and GSM9196 through GSM9207; whole genome arrays, GSM8942 through GSM8977 and GSM9208 through GSM9219. GEO accessions for pilot and whole genome array analysis files: GSE601, GSE636–639. GenBank accession numbers can be found on table S6.

Supporting Online Material

www.sciencemag.org/cgi/content/full/302/5646/842/DC1
Materials and Methods
SOM Text
Figs. S1 to S27
Tables S1 to S6
References

23 June 2003; accepted 10 September 2003

REPORTS

Direct Atom-Resolved Imaging of Oxides and Their Grain Boundaries

Zaoli Zhang, Wilfried Sigle, Fritz Phillipp, Manfred Rühle*

Using high-resolution transmission electron microscopy, we obtained structure images of strontium titanate ($SrTiO_3$) with a clearly resolved oxygen sublattice along different crystallographic directions in the bulk lattice and for a $\Sigma 3$ tilt grain boundary. Comparison with image simulations showed that the grain boundary contains oxygen vacancies. Measurements of atom displacements near the grain boundary revealed close correspondence with theoretical calculations.

Direct imaging of atoms in oxides, especially in defect regions, is of considerable importance for understanding the properties of materials. Because oxygen has a high electronegativity, its presence can have a marked effect on these properties. This can occur through the binding of free electrons, which leads to a loss of electrical conductivity (e.g., in metal oxides) through the presence of vacancies in the oxygen sublattice, which in turn can give rise to considerable ionic conductivity (as in electroceramic materials). Oxygen has a strong effect on mechanical properties, which are directly linked with the strong interatomic forces. But oxygen can also have an indirect influence: Most properties of materials are sensitively influenced by the presence of crystal defects, and the oxygen concentration at such defects can drastically deviate from the average bulk value, even in thermal equilibrium. This is one of the reasons why grain boundaries (GBs) in electroceramics can be electrically charged and form a double Schottky barrier against the movement of ions (*1*, *2*), which is of ultimate importance in polycrystalline electroceramic materials. Recently, it was shown that dislocations in $SrTiO_3$ can be oxygen deficient (*3*) and that this can lead to a change

Max-Planck-Institut für Metallforschung, Heisenbergstraße 3, D-70569 Stuttgart, Germany.

*To whom correspondence should be addressed. E-mail: ruehle@mf.mpg.de